UNIVERSITY OF BRIGHTON

MODULE TUTOR: - Dr Gulden Uchyigit

MODULE CODE: - CI512 Intelligent Systems Sem-1

STUDENT ID: - 21839630

SUBMISSION DATE: - 18/08/2023

Table of Contents

[Introduction 2](#_Toc143245798)

[Aim of the project 2](#_Toc143245799)

[About the Dataset 2](#_Toc143245800)

[Solution 2](#_Toc143245801)

[Data visualization And Analysis 2](#_Toc143245802)

[Information of the features 2](#_Toc143245803)

[Checking Null values and missing values 3](#_Toc143245804)

[Correlation Heatmap 4](#_Toc143245805)

[Death Event Vs Age 4](#_Toc143245806)

[Smoker vs death events 5](#_Toc143245807)

[Death events Vs sex 6](#_Toc143245808)

[Death events vs anaemia 7](#_Toc143245809)

[Death events vs Diabetes 8](#_Toc143245810)

[Ejection fraction VS Death event 9](#_Toc143245811)

[Data Preparation and splitting of data. 10](#_Toc143245812)

[Model Implementation And evaluation 11](#_Toc143245813)

[**Logistic regression** 11](#_Toc143245814)

[MODEL IMPLEMENTATION 11](#_Toc143245815)

[Confusion Matix 12](#_Toc143245816)

[Classification Report 13](#_Toc143245817)

[XGBoost 13](#_Toc143245818)

[Model Implementation 13](#_Toc143245819)

[Confusion Matix 14](#_Toc143245820)

[Classification Report 15](#_Toc143245821)

[**K-Nearest Neighbors (KNN)** 15](#_Toc143245822)

[Model Implementation 15](#_Toc143245823)

[Confusion Matix 16](#_Toc143245824)

[Classification Report 16](#_Toc143245825)

[Comparison of FN and FP 16](#_Toc143245826)

[Conclusion 17](#_Toc143245827)

# Introduction

## Aim of the project

The aim of this project is to develop a predictive model using artificial intelligence techniques to determine whether a patient is at risk of experiencing heart failure. By data analysis and machine learning, the project aims to enhance medical diagnostics.

## About the Dataset

Heart failure, a prevalent outcome of cardiovascular diseases, is predicted through a dataset encompassing 12 predictive attributes related to mortality by heart failure. The problem of cardiovascular diseases largely relies on behavioral risks like tobacco usage, poor diet, inactivity, and excessive alcohol consumption, necessitating population-level interventions. Identification of the diseases and treatment of cardiovascular concerns, particularly in high-risk individuals with hypertension, diabetes, or hyperlipidemia, can be facilitated by machine learning models.

## Solution

The project will employ a diverse set of models to predict the risk of heart failure prioritizing accuracy and precision. By minimizing both false negatives and false positives, the goal is to ensure optimal performance in identifying potential cases. These models will undergo rigorous evaluation and comparison, with the ultimate selection being based on a combination of high accuracy and minimal false negatives and false positives. This approach will result in the identification of the most effective model for accurately predicting heart failure while minimizing the risk of misdiagnoses.

# Data visualization And Analysis

## Information of the features

Upon inspecting the data set, it was found that information from 299 patients is encompassed by 13 distinct attributes. Within these attributes, 7 are Numeric in nature, representing quantities of specific features present in patients. The remaining 6 attributes were in Boolean format, indicating the presence or absence of conditions for each patient. This comprehensive dataset structure provides a rich foundation for further analysis, allowing for exploration of patient characteristics and their potential impact on heart failure prediction.

A white paper with black text

Description automatically generated

## Checking Null values and missing values

Using functions such as df.isnull, df.info, and df.describe, which are used for extracting data insights, inspecting null values, and checking dataset characteristics, a thorough evaluation was conducted. This highlighted that there was absence of any null or missing values within the dataset. With this information we can start more detailed and comprehensive analysis, enabling confident progress in exploring intricate patterns and relationships within the data.

A screenshot of a computer

Description automatically generated

## Correlation Heatmap

Correlation maps were used for checking the relationship between the features. This Provides the overview of the data set. Positive correlations were depicted in red, while negative correlations were represented in blue. Our correlation analysis revealed significant positive links between age and death events, death events and serum creatinine levels, as well as smoking and gender. Conversely, notable negative associations surfaced, including those between death events and time, age and time, and death events and ejection fraction. These findings illuminate pivotal connections within the data, guiding subsequent analytical directions and model refinement.

A graph with red squares

Description automatically generated

## Death Event Vs Age

A comparative histogram analysis was employed to assess the relationship between age and death events, differentiating survivors from non-survivors. The analysis aimed to identify the age range with the highest death occurrences with survival statistics. Notably, the dataset contains 203 survivors and records 96 instances of death. Upon scrutinizing the results, a distinct pattern emerged most deaths clustered within the age range of 40 to 70, with a pronounced peak at age 60. These observations accentuate the age-specific vulnerability to heart failure and provide critical insights for potential intervention strategies.

A graph of a number of people

Description automatically generated

## Smoker vs death events

To analyze the relationship between smoking status and death events, a bar graph was employed to illustrate the distribution of smokers within the dataset across different categories: smokers who survived, smokers who died, non-smokers who survived, and non-smokers who died. The results revealed a higher count of deaths among individuals who smoked (137), compared to non-smokers who died (66) and smokers who survived (66). A notably lower count of survivors who were non-smokers (30) further underscores the impact of smoking on heart failure outcomes, highlighting the importance of lifestyle factors in patient prognosis.

A graph of smoking status

Description automatically generated

Indeed, due to the presence of heart attack cases in the dataset, the observed ratio of smokers to non-smokers may not accurately represent the general population. The dataset's inherent bias towards individuals with heart conditions could lead to misleading conclusions about the prevalence of heart failure among smokers and non-smokers. Therefore, while the bar graph provides insights, it's essential to interpret the findings cautiously and consider the potential influence of dataset composition on the observed patterns. Further analysis and potentially a control group could be necessary to ascertain a more comprehensive understanding of the association between smoking and heart failure incidence.

## Death events Vs sex

the relationship between death events and gender, it is evident that the number of male deaths attributed to heart failure surpasses that of females. Specifically, male deaths total 132, while female deaths amount to 71—half the male count. This trend is mirrored in the survivor category, where male survivors total 62 and female survivors total 34. This gender-based analysis suggests a higher prevalence of heart failure-related deaths among males compared to females within the dataset, highlighting potential gender-specific differences in heart failure susceptibility and outcomes.

A graph of a number of people

Description automatically generated

## Death events vs anaemia

Upon examining the relationship between death events and the presence of anemia, a notable trend emerges individuals with anemia exhibit a higher likelihood of experiencing a death event compared to those without anemia. Specifically, slightly over half of the individuals with anemia succumbed to death, in contrast to a little over a third of those without anemia. This observation underscores the potential impact of anemia as a contributing factor to heart failure-related mortality, implying the need for further investigation into the association between anemia and adverse cardiac outcomes.

A bar chart with numbers and text

Description automatically generated

## Death events vs Diabetes

Analyzing the distribution of death events in relation to diabetes, a minimal distinction is evident between patients who died and had diabetes, and patients who are alive and have diabetes. The ratios of these two groups are nearly identical, indicating that the presence of diabetes alone might not significantly differentiate between patients who experienced death events and those who survived. This finding underscores the complex nature of heart failure prediction, suggesting that factors beyond diabetes alone may contribute to adverse outcomes. The exploration of combined risk factors is warranted to gain a more comprehensive understanding of heart failure dynamics.

A screenshot of a graph

Description automatically generated

## Ejection fraction VS Death event

The relationship between ejection fraction and death events reveals intriguing patterns. When ejection fraction ranges from 25 to 55, death events are relatively low. However, the risk of death events is elevated when ejection fraction is below 30 or exceeds 60. Notably, death events are nearly equal when ejection fraction is precisely at 25. This implies that patients with ejection fractions between 25 and 55 have better survival odds, while those with values outside this range face higher risks. This insight suggests ejection fraction as a critical indicator, potentially guiding clinical interventions for improved heart failure prognosis.

A graph of a number of people

Description automatically generated

# Data Preparation and splitting of data.

The SelectKBest feature selection method available in the sklearn library to determine the most influential features for predicting the 'DEATH\_EVENT' in a dataset. I have divided the data by splitting the data into two component the features (X) and the target variable (y), where 'DEATH\_EVENT' is the variable to be predicted. I have implemented the SelectKBest class with the chi2 scoring function, which is suitable for categorical target variables. The parameter 'k' is set to determine the number of features to select. By applying the fit\_transform method of the SelectKBest object to the feature matrix X and the target y, a new feature matrix named X\_k\_best is obtained. This new matrix contains only the selected features which give most significant for predicting the death event outcomes. Then printed out the best features and they were, age', 'creatinine\_phosphokinase', 'ejection\_fraction', 'platelets', 'time.

A screenshot of a computer

Description automatically generated

# Model Implementation And evaluation

The choice of a suitable model depends on various factors, encompassing the nature and size of the dataset. In the context of heart failure prediction, given the specific characteristics of the data and its volume, I have opted to implement three distinct algorithms: logistic regression, K-nearest neighbors (KNN), and XGBoost. While implementing the model I have also given unique id for each of them.

For evaluation purposes the confusion matrix, classification report, and comparison of FN and FP are implemented. With the unique ID given to the model they don’t give wrong output.

## **Logistic regression**

Logistic regression is a reliable choice when dealing with binary classification tasks, like predicting heart failure. Its simplicity and interpretability make it suitable for initial analyses, aiding in understanding the relationships between variables.

### MODEL IMPLEMENTATION

I started by importing the LogisticRegression class from the sklearn.linear\_model module, which is used to build a logistic regression model. Then I initialized and trained the model by giving an instance of the LogisticRegression class named lr. The parameter random\_state is set to ensure duplicability of the results. The model is then trained on the training data using the fit method, where X\_train represents the feature matrix of the training data, and y\_train is the corresponding target variable. For making the model to predict The trained logistic regression model is employed to predict outcomes for the test data. The predict method is utilized, with X\_test being the feature matrix of the test data. The predictions are stored in the variable predicted1.

A screenshot of a computer code

Description automatically generated

### Confusion Matix

A computer screen shot of a computer code

Description automatically generated

The confusion matric had unique id for logistic regression which is cm\_lr. The confusion matrix predicted that the logistic regression model had 20 true negative, 2 false positive , 2 false negative and 6 true positives

### Classification Report

A screenshot of a computer

Description automatically generated

## XGBoost

XGBoost, known for its ensemble nature and robustness, is effective at handling complex relationships within data. Its boosting approach builds a strong predictive model by combining the strengths of numerous weak learners, which can enhance predictive accuracy.

### Model Implementation

I have imported modules which includes XGBClassifier from xgboost.sklearn, and RandomizedSearchCV and GridSearchCV from sklearn.model\_selection. Defining of hyperparameter grid is done A dictionary named parameters is defined, containing various hyperparameter values for the XGBoost classifier. Parameters include learning rate, max depth, min child weight, subsample, colsample bytree, number of estimators, gamma, reg alpha, and reg lambda and The RandomizedSearchCV object xgb\_grid is instantiated. This object takes the model, parameters, cross-validation (cv) value, and the number of CPU cores (n\_jobs) for parallel processing. The xgb\_grid object is fitted to the training data (X\_train, y\_train) using the randomized search. This process explores various hyperparameter combinations, evaluating their performance using cross-validation.

A screenshot of a computer program

Description automatically generated

After tuning the XGBoost classifier using the randomized search, you can apply this optimized model to make predictions on new data. The predict method of the xgb\_grid object is utilized for this purpose. The labels are created for The predictions for the test dataset (X\_test) are generated using the trained and tuned XGBoost classifier, and the predicted labels are stored in the variable predicted2.

### Confusion Matix

A computer screen shot of a computer code

Description automatically generated

The confusion matric had unique id for XGBoost which is cm\_xgb. The confusion matrix predicted that the XGBoost model had 21 true negative, 1 false positive, 2 false negative and 6 true positives

### Classification Report

A screenshot of a test results

Description automatically generated

## **K-Nearest Neighbors (KNN)**

KNN, on the other hand, can be advantageous when there's a likelihood of local patterns and interactions in the data. This algorithm calculates predictions based on the majority class among the k-nearest neighbors of a data point, making it sensitive to localized structures.

### Model Implementation

The necessary modules were imported for implementing the knn module, which are The code imports the KNeighborsClassifier class from sklearn.neighbors and the GridSearchCV class from sklearn.model\_selection. For initializing the knn classifier an instance of the KNeighborsClassifier named knn is created. For Hyperparameter Grid a dictionary named k\_values are defined, containing a range of n\_neighbors values from 1 to 9.

A screenshot of a computer code

Description automatically generated

For the Grid Search for Hyperparameter The GridSearchCV object grid is instantiated. This object takes the knn classifier, k\_values dictionary, cross-validation (cv) value, scoring metric (scoring), and the number of CPU cores (n\_jobs) for parallel processing. The trained and tuned KNN model (grid) is utilized to predict labels for the test dataset (X\_test) using the predict method. The predicted labels are stored in the variable predicted3.

The best parameters were 6

### Confusion Matix

A screen shot of a computer code

Description automatically generated

The confusion matric had unique id for KNN which is knn\_xgb. The confusion matrix predicted that the KNN model had 19 true negative, 3 false positive, 8 false negative, 0 true positive

### Classification Report

A screenshot of a computer

Description automatically generated

## Comparison of FN and FP

The bar graph shows that the FP and FN of KNN is higher compared to Xgboost and logistic regression. On comparison between the logistic regression and xgboost, the FP are higher for logistic regression than xgboost. But the FN for both logistic regression and XGBoost are equal.

A graph of different colored squares

Description automatically generated

# Conclusion

In conclusion, after thorough analysis and model evaluation, it is evident that the XGBoost algorithm emerges as the optimal choice for heart failure prediction. With an impressive accuracy rate of 90 percent, the XGBoost model showcases superior performance. Furthermore, this model demonstrates commendable results in terms of minimizing both false negatives and false positives, indicative of its robustness in capturing the complexities of heart failure prediction.

In contrast, the alternative models show comparatively lower accuracies. Logistic regression attains an accuracy of 87 percent, indicating solid predictive capabilities. On the other hand, K-nearest neighbors (KNN) achieves a modest accuracy of 63 percent, suggesting that it may not fully capture the intricate patterns present in the heart failure data.

Given its superior accuracy and balanced handling of false negatives and false positives, the XGBoost model emerges as the most reliable and effective option for heart failure prediction in this analysis.

# Reference

L;, C.D. (no date) *An enhanced random forests approach to predict heart failure from small imbalanced gene expression data*, *IEEE/ACM transactions on computational biology and bioinformatics*. Available at: https://pubmed.ncbi.nlm.nih.gov/33259306/ (Accessed: 18 August 2023).

*Predicting heart disease using machine learning* (2019) *YouTube*. Available at: https://youtu.be/75OJvlhFUMY (Accessed: 18 August 2023).

Pritishmishra (2021) *Logistic regression on heart disease dataset*, *Kaggle*. Available at: https://www.kaggle.com/code/pritishmishra/logistic-regression-on-heart-disease-dataset (Accessed: 18 August 2023).

Ulianova, S. (2019) *Cardiovascular disease dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset (Accessed: 18 August 2023).

Wang1, J. (2021) *IOPscience*, *Journal of Physics: Conference Series*. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/2031/1/012068 (Accessed: 18 August 2023).

*Welcome to the UC Irvine Machine Learning Repository* (no date) *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ (Accessed: 18 August 2023).